Sun Fire X4500 und Solaris ZFS - Neue Lösungen für große Datenmengen

Joachim Krebs
Senior Systemberater
Sun Microsystems GmbH, Stuttgart

Black Forest Grid

computer center university of freiburg

3rd Black Forest Grid Workshop

April 19-20 2007

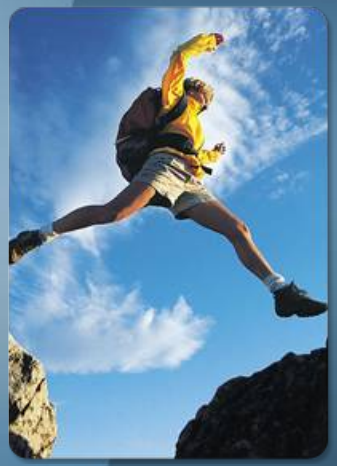
Agenda

- Introduction
- The Sun Fire X4500 Server and Use Cases
- Solaris 10 ZFS
- Summary

Massive growth in Data creates specific challenges

How to afford the storage needs of the future

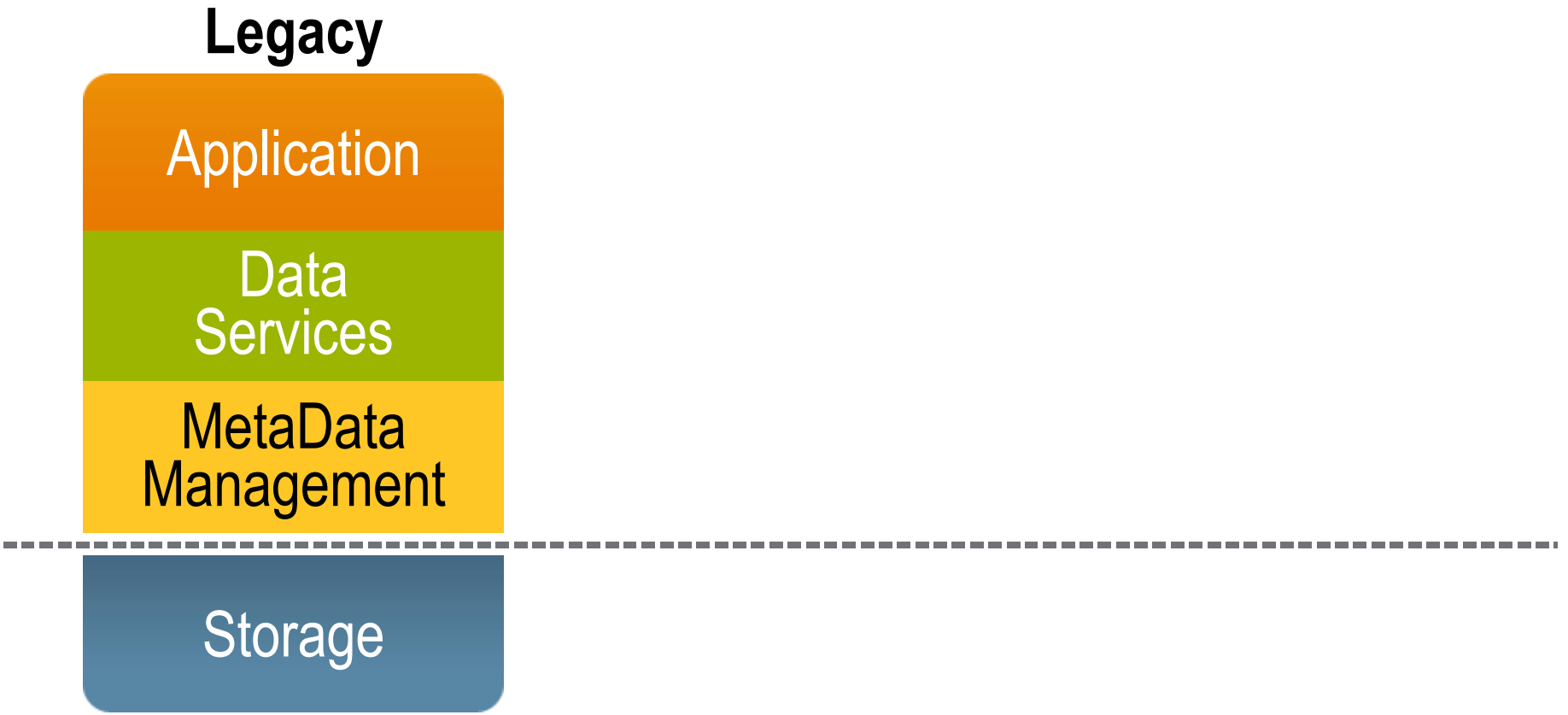
> 390 Gigabytes of data created each second
> Data growing at rate of 50% per year
> Cost of storing data dominates decision making
> Rising costs of real estate, power and cooling



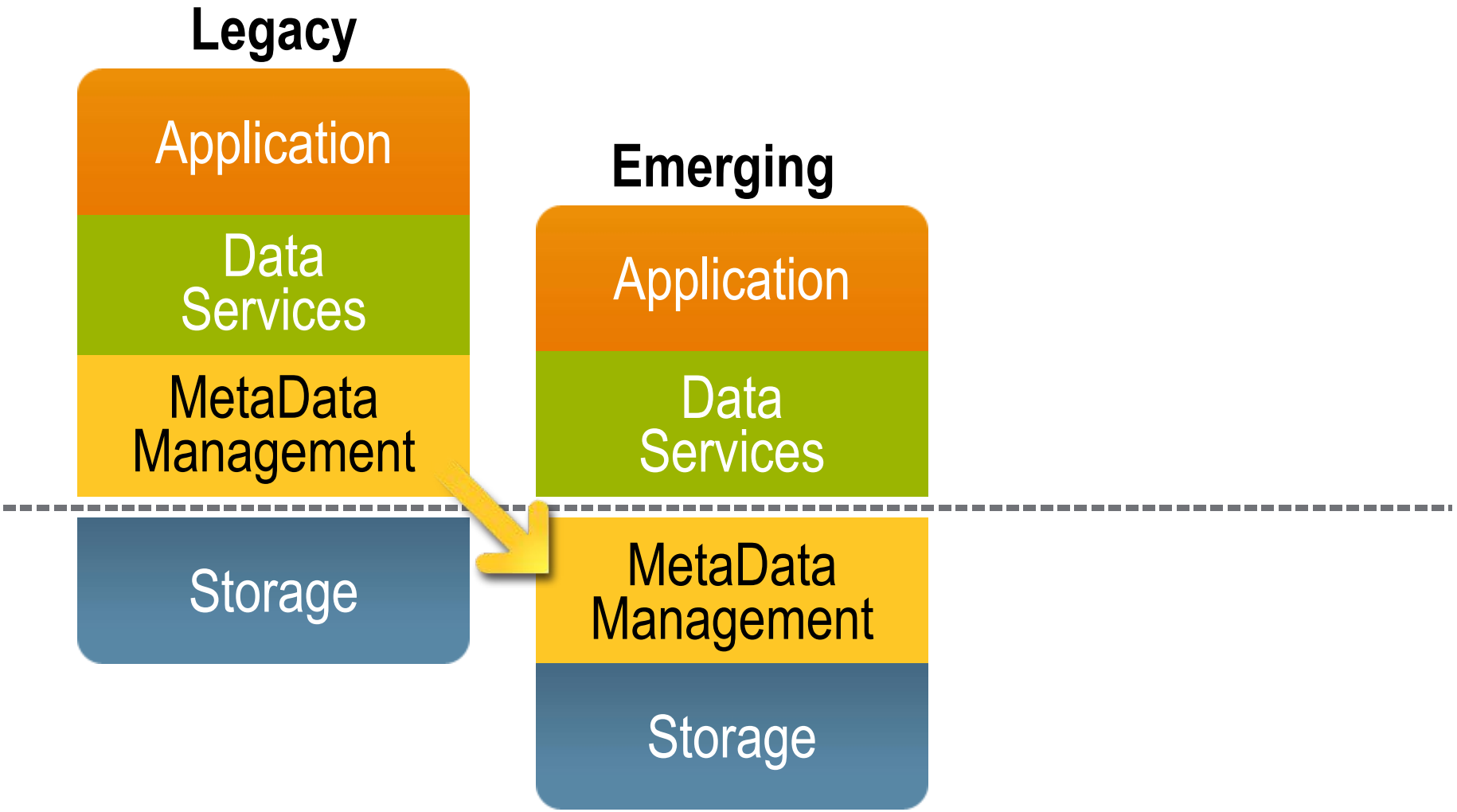
How to access and process data quickly and cost effectively

> Need low cost way to stream data over the network at high speeds
> Need to feed applications with large amounts of data very quickly

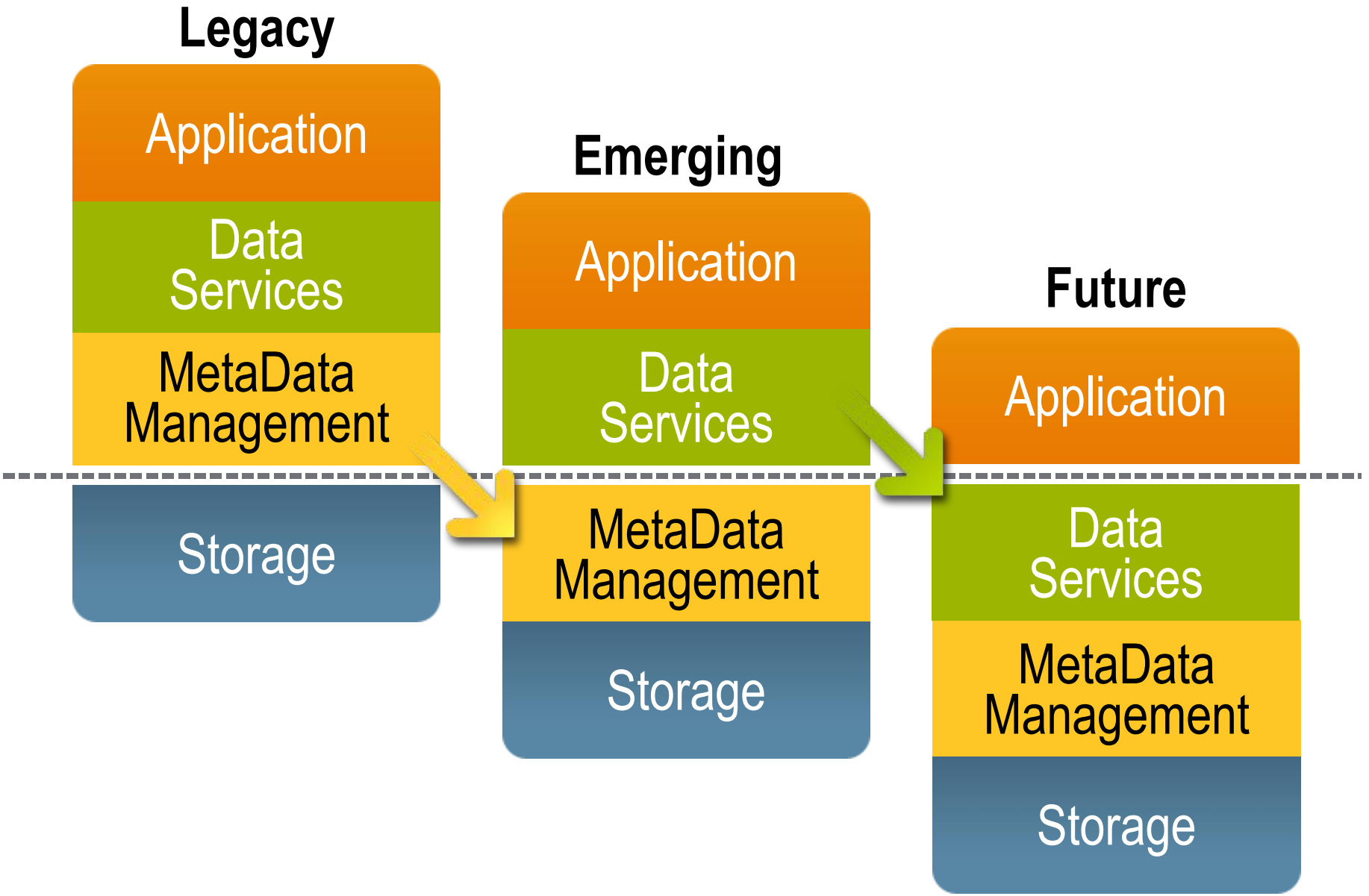
Delivering New Storage Paradigms



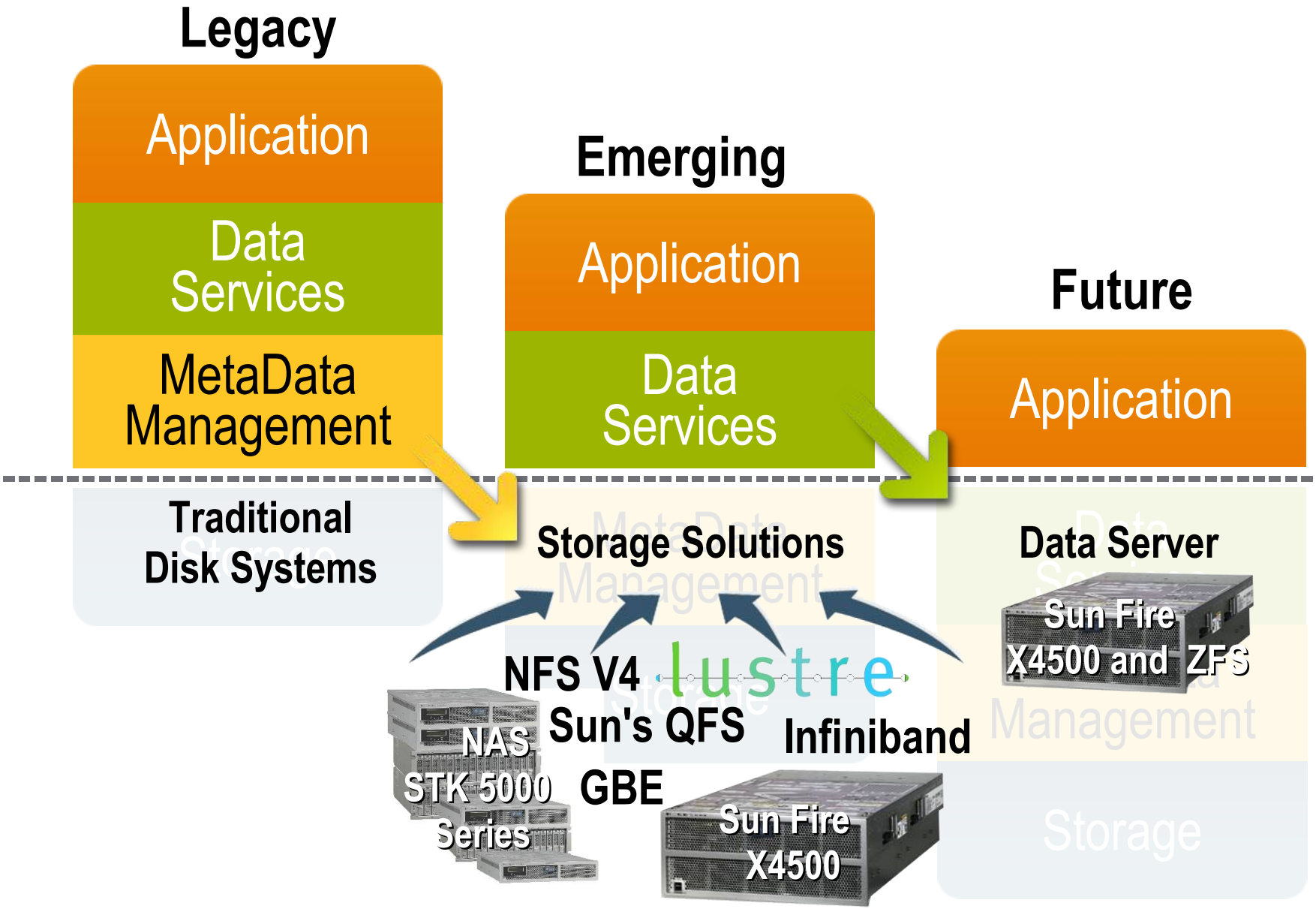
Delivering New Storage Paradigms



Delivering New Storage Paradigms



Delivering New Storage Paradigms



Thumper
The only Hybrid Data Server

Sun Fire X4500 Server

Agenda

- Introduction
- The Sun Fire X4500 Server and Use Cases
- Solaris 10 ZFS
- Summary

Sun Fire X4500 Server

Integrated x86 compute power, massive storage capacity and high data throughput

HPC, Data Warehouse/Business Intelligence, Digital Media, Digital Surveillance, VTL

post GA

Compute

- 2 x Dual Core Opteron processors
- 16GB Memory

Storage

- 48 Sata disks
- Up to 24TB raw capacity

I/O

- Very high throughput
- 2x PCI-X slots
- 4 GigE

Availability

- Hot-swap/plug power, fans, disks

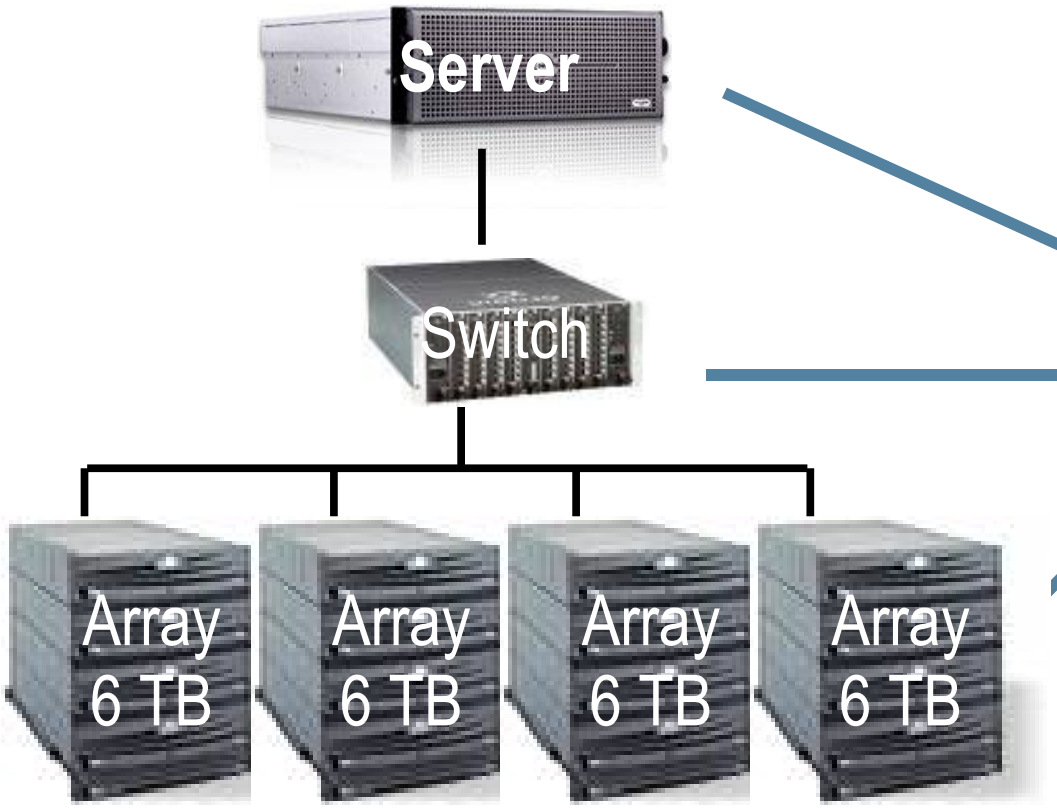
Management

- Same management as other Galaxy servers

Solaris(TM) ZFS

- Ground-breaking file system performance

New Architectural Concept: Sun Fire X4500



Single Integrated System

Server + SAN-Switch+ Storage Arrays

Sun Fire X4500 Server

Sun Fire X4500 Server

- **A Member of the Galaxy Family of Servers**
  2 Opteron Dual-Core processors
  16GB Memory (2GB Dimms)
- **With on Board High Density SATA**
  48 direct attached hot-plug SATA drives
  24TB in 4 RU
- **Delivering Incredible Throughput**
  1 GB/s Disk to Network Design Goal
  4 GigE ports and 2 PCI-X slots
- **And Enterprise Class Server RAS**
  Redundant Fans, Power Supplies & Hard Disk & ILOM

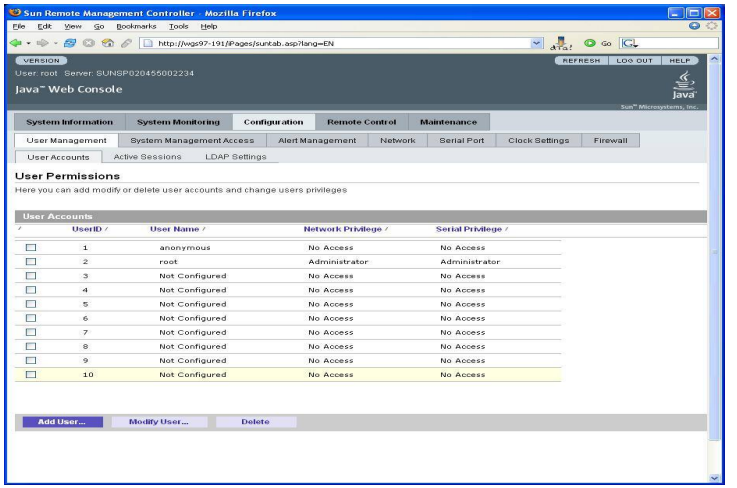
n+x

Sun's x64 System Portfolio

ILOM – Integrated Lights Out Manager

Service Processor Technology for Sun Fire Systems

- Same standard Lights Out Management for Sun Fire family systems
- Provides full local or remote access for setup, maintenance and on-going monitoring/management of a single system
- Full remote KVM functionality
  - > Including remote media support
- Browser-based UI and full CLI
- Access via Management Ethernet port,
- Serial port or Host OS (with suitable driver)
- Standards supported include LDAP, SSH 2.0, SNMP v1, v2c, v3, IMPI 2.0, DMTF 'SMASH' CLI

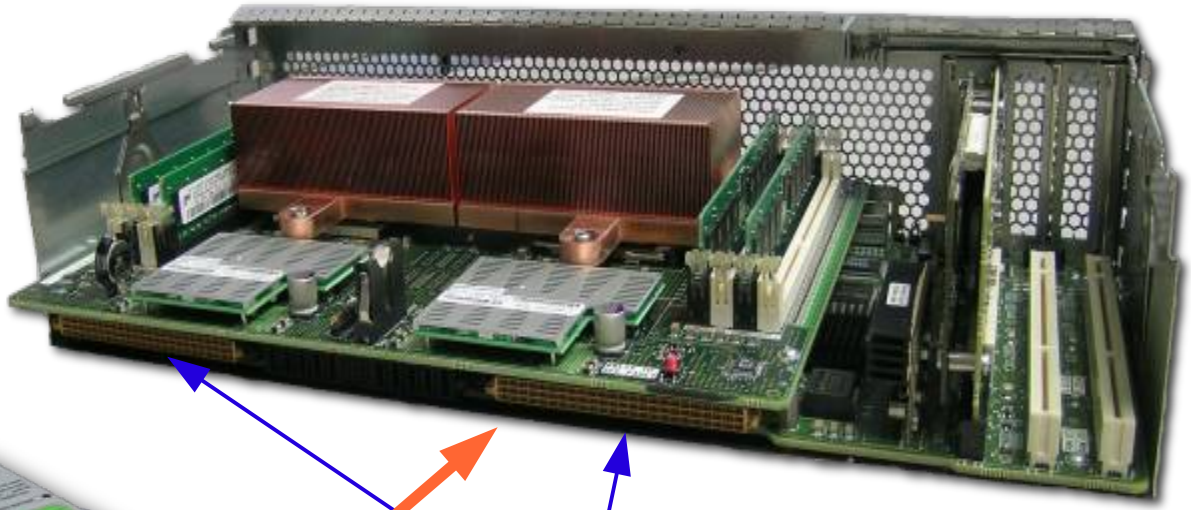


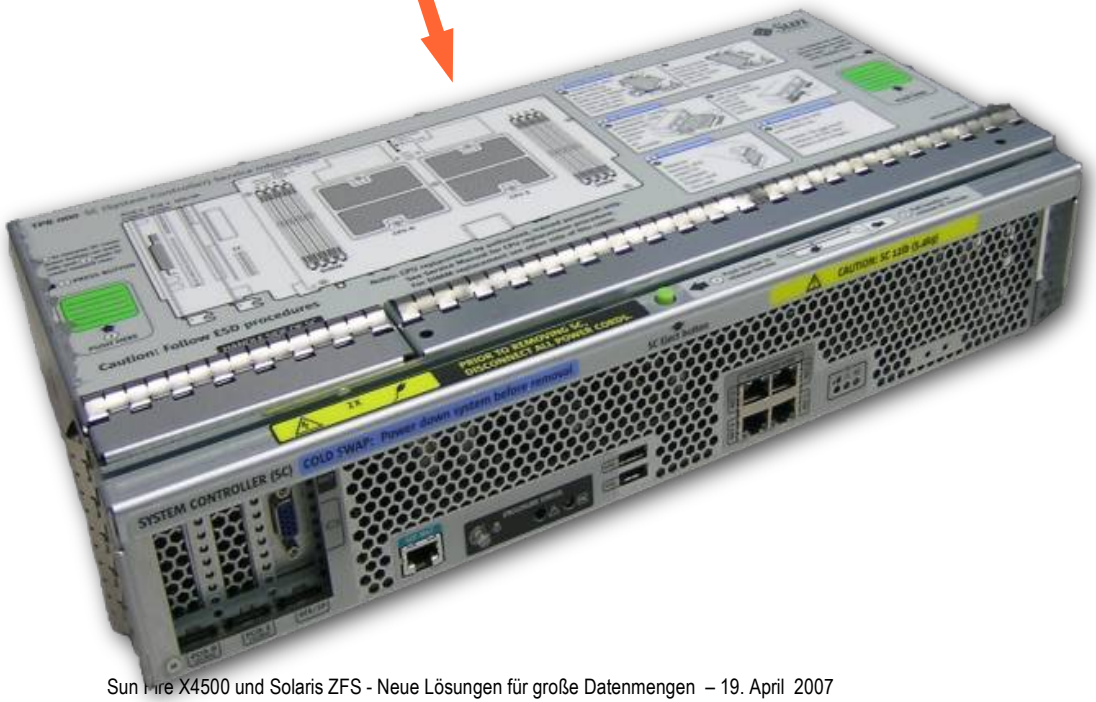
Sun Fire X4500 Server

Server Module

Mother board is split in two halves: Processor board and I/O board

Back View





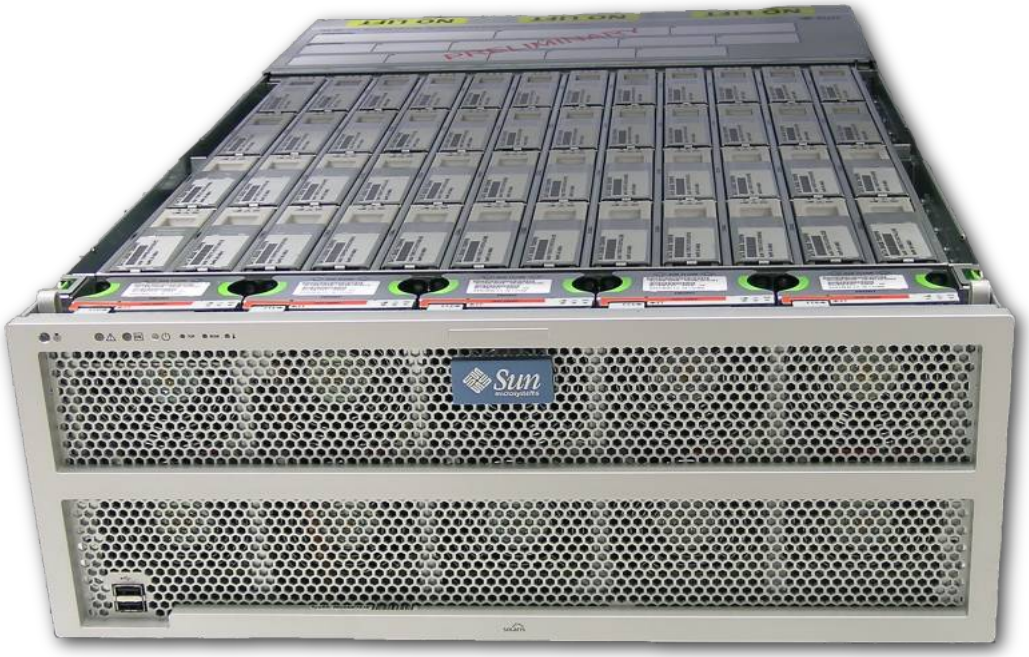
Disk Backplane Connectors

PCI-X Expansion Slots

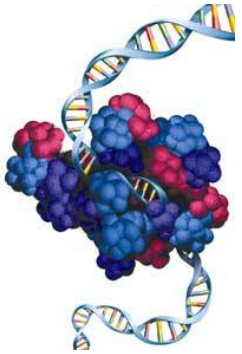
- 2 PCI-X MD2 expansion slots
- Tooless service
- Uses 3.3 Volt signaling bus
- 32 or 64 bit
- Each utilizes a separate PCI-X bus
- Maximum speed 133 MHz/64 bit
- Maximum throughput 1.06 GBps (8.5 Gbps)
- Supports low profile MD2 PCI or PCI-X expansion cards

Storage detail

- Marvell 88SX6081 8-port Serial-ATA 2.0 (SATA II) Storage Controllers
- Hitachi Deskstar 7K250/7K500 3.5-inch SATA II disk drives
- Choices of 500 GB 7200-RPM hard disk drives
- Cable-less disk backplane



Use Cases in High Performance and Grid Computing



Challenge

> Soaring storage costs, compute node utilization and physical space limitations

Sun Fire X4500 Server Benefit

> The only solution that combines low-cost, high density storage with throughput rates that scale as compute nodes are added to meet the needs of modern HPC applications.

~1/2 Cost

1/5 Space

Compared to alternatives

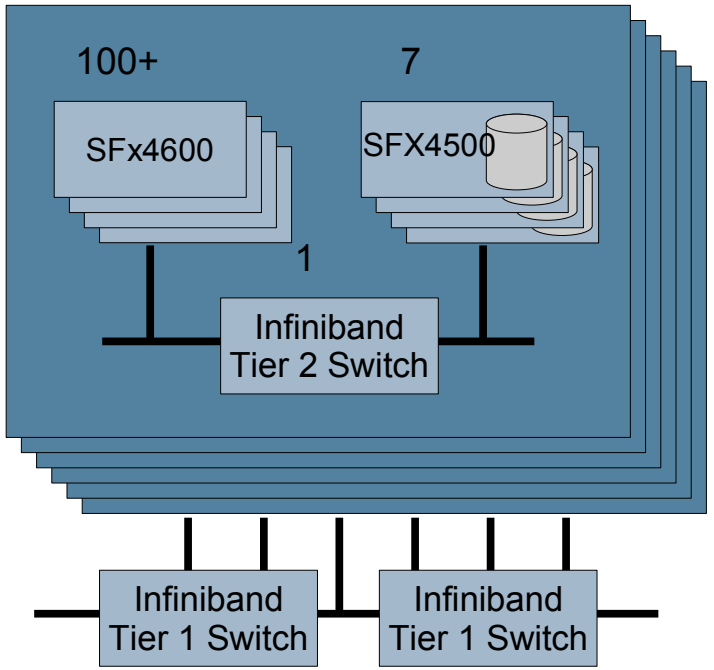
Use Case: Tsubame Supercomputer

Tokyo Institute of Technology (TiTech)

Fastest Supercomputer Outside the U.S.

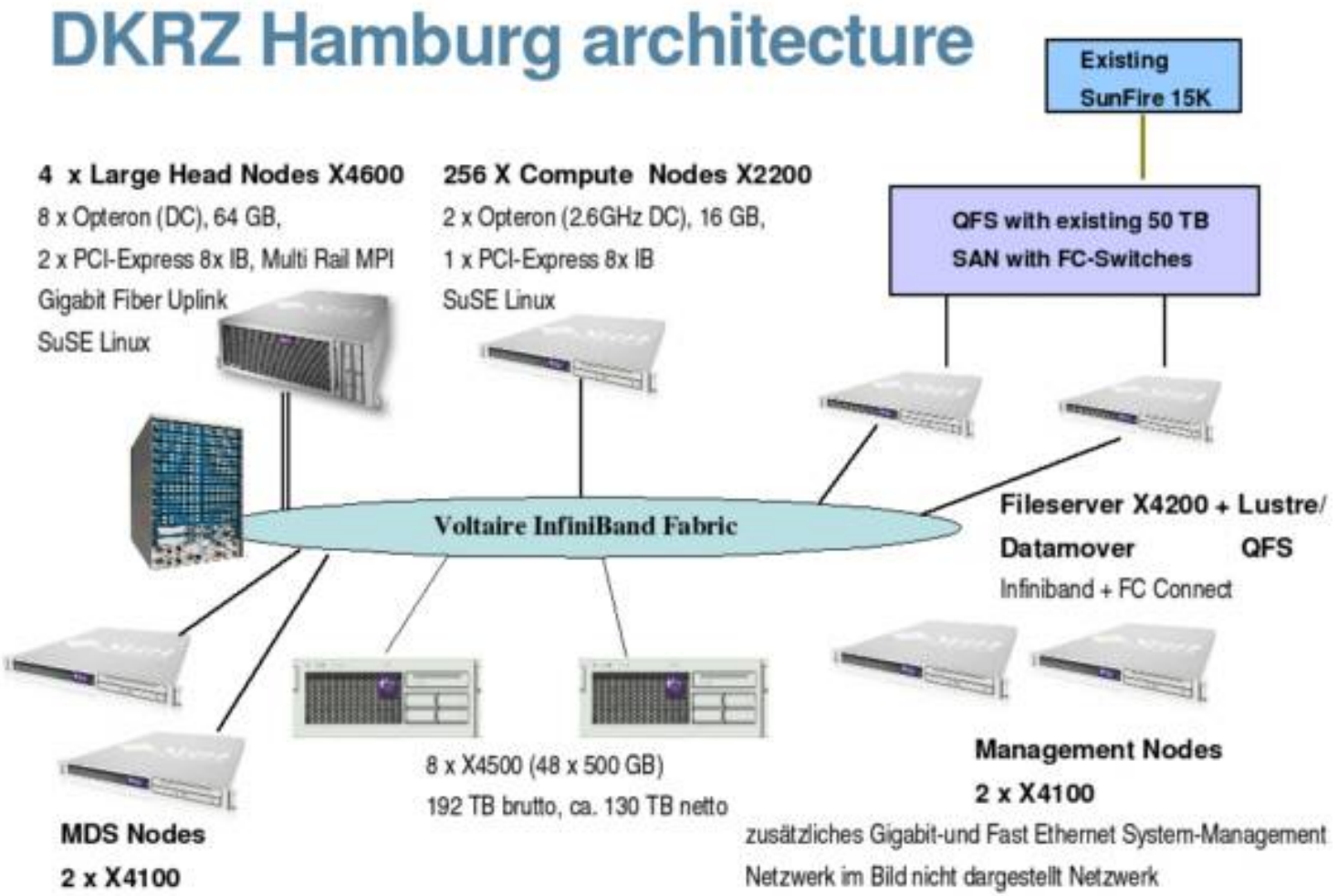
Tsubame By The Numbers

36.36 TFLOPS in 31 Days



6	linked sub-clusters
8	Voltaire Infiniband switches
42	**Sun Fire X4500 Data Servers**
80+	racks
360	Clearspeed FP accelerators
655	**Sun Fire X4600 Servers**
10480	**Opteron cores**
21	TeraBytes RAM
1	**PetaByte storage**
	N1 Grid Engine
	N1 System Manager
	Lustre parallel file system

Use case: Linux Cluster - Deutsche Klima Rechenzentrum



http://www.dkrz.de/dkrz/about/hardware/linux_cluster

Agenda

- Introduction
- The Sun Fire X4500 Server and Use Cases
- Solaris 10 ZFS
- Summary

Solaris ZFS

End-to-end data integrity

Immense data capacity

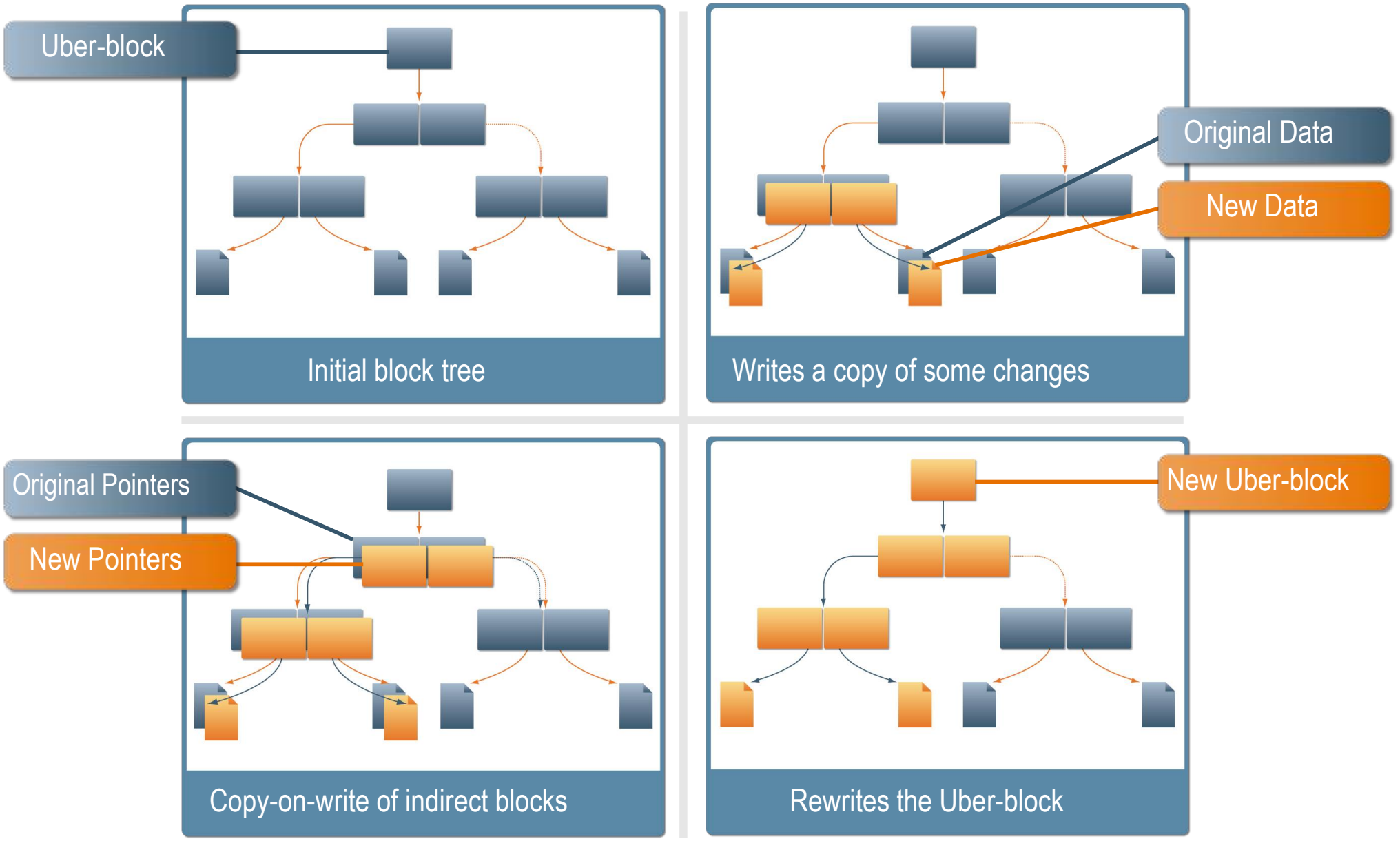
Easier administration lowers costs

Huge performance gains

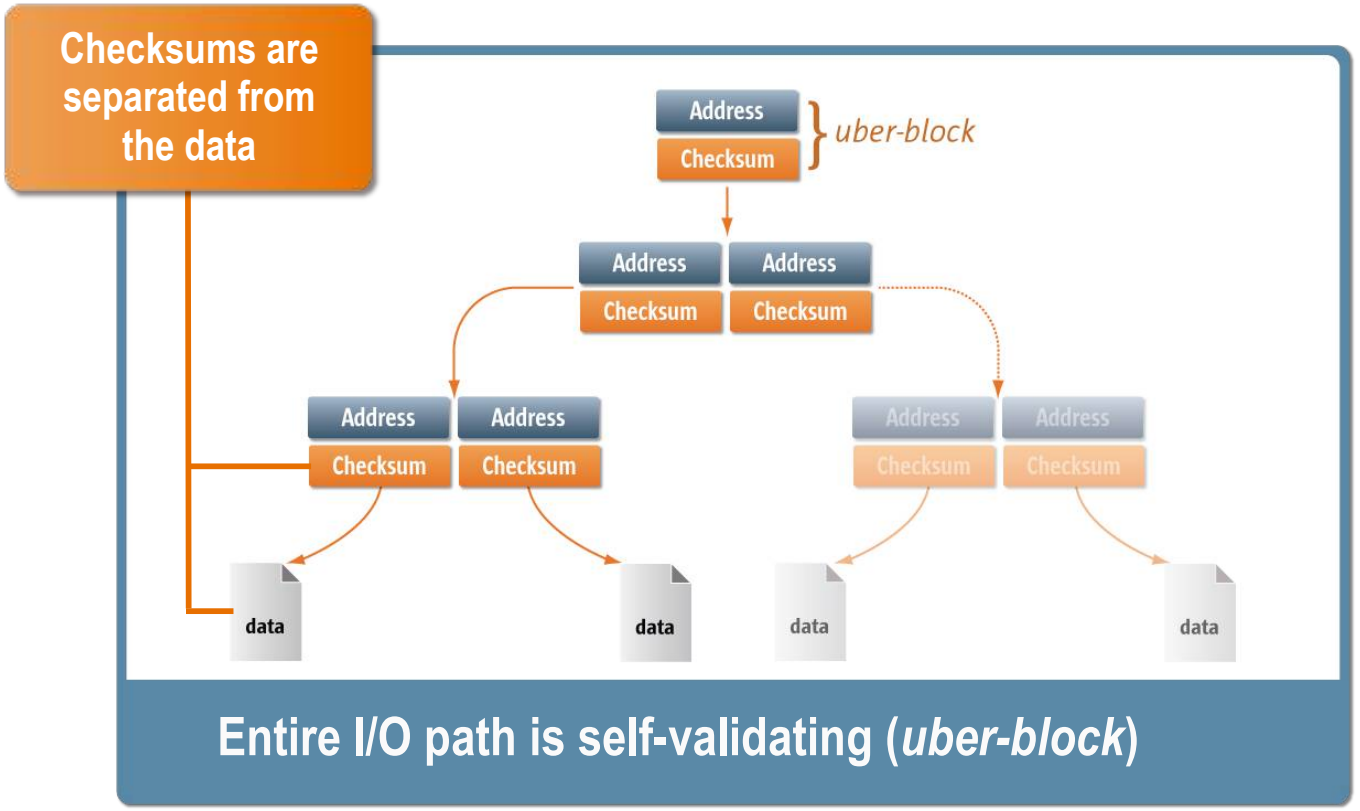
ZFS Data Integrity Model

- Copy-on-write, transactional design
- Everything is checksummed
- RAID-Z and RAID-Z2 (Solaris 10 11/06) protection
- Hot Spare
- Disk Scrubbing

Copy-on-Write and Transactional



End-to-End Checksums



Prevents:

- Silent data corruption
- Panics from corrupted metadata
- Phantom writes
- Misdirected reads and writes
- DMA parity errors
- Errors from driver bugs
- Accidental overwrites

Self-Healing Data

ZFS can detect bad data using checksums and “heal” the data using its mirrored copy.

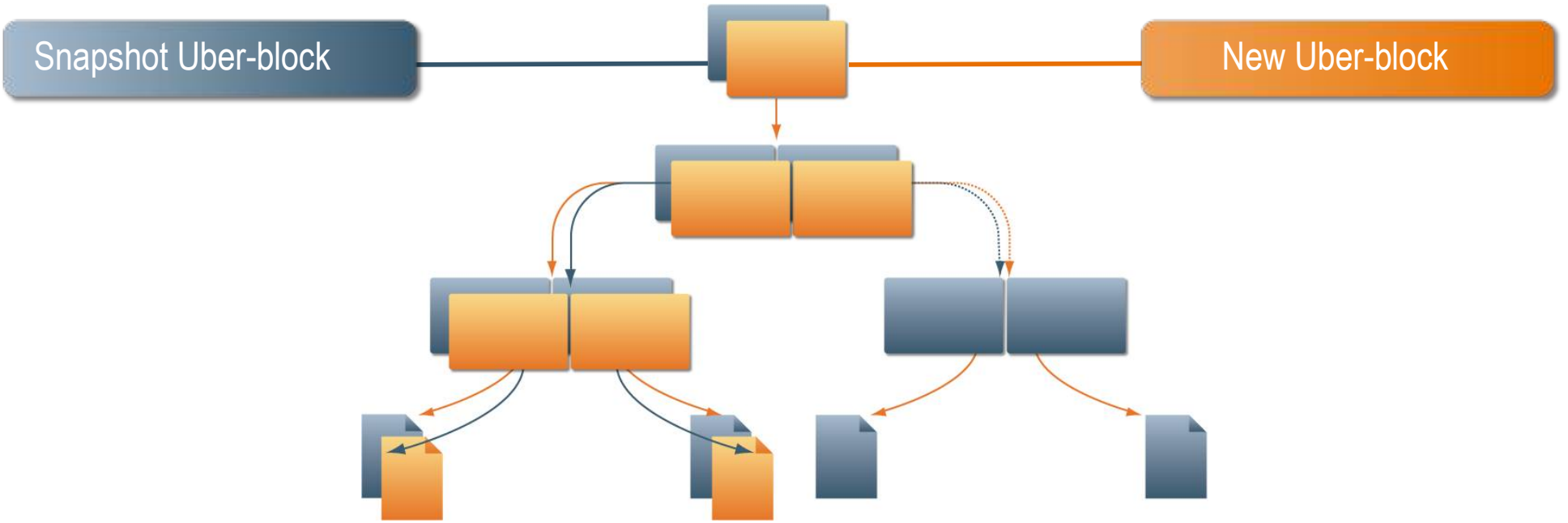
Detects Bad Data

Gets Good Data from Mirror

“Heals” Bad Copy and Sends Back to Application

ZFS Snapshots

- Provide a read-only point-in-time copy of filesystem
- Copy-on-write makes them essentially “free”
- Very space efficient – only changes are tracked
- And instantaneous – simply retains the old structure

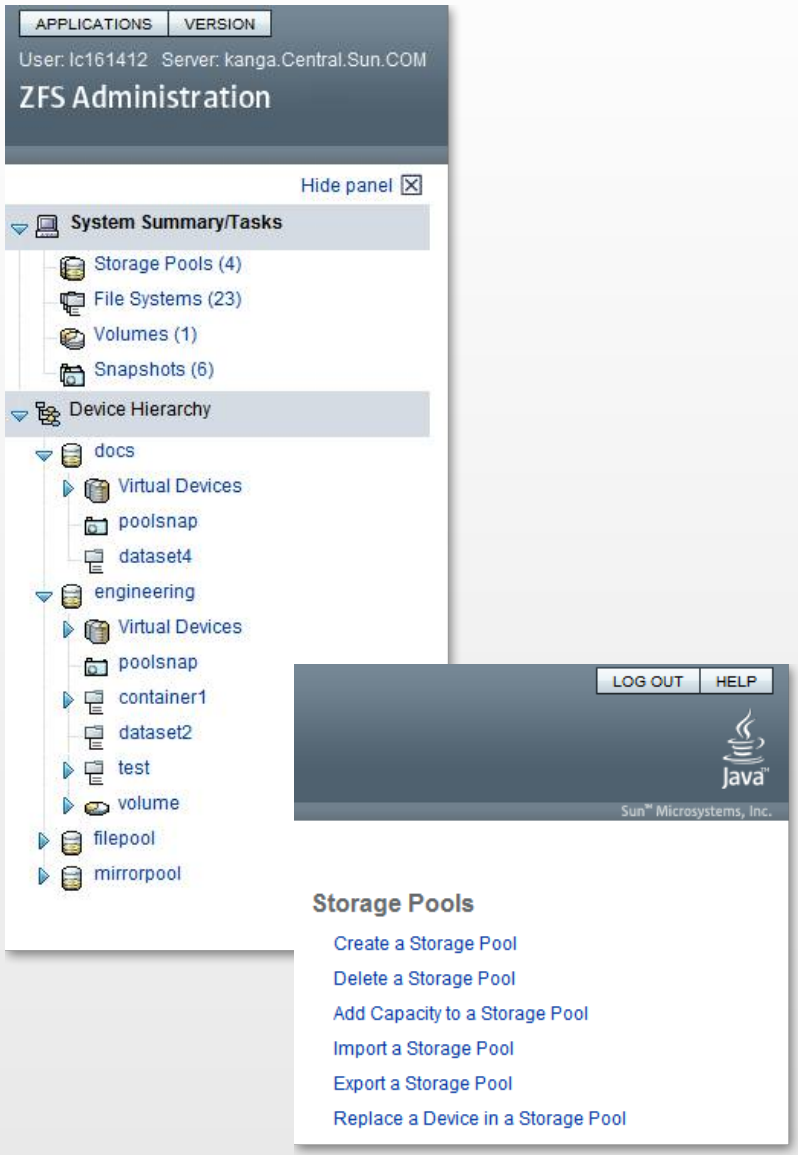


Easier Administration

- Pooled Storage Design makes for Easier Administration

No need for a Volume Manager!

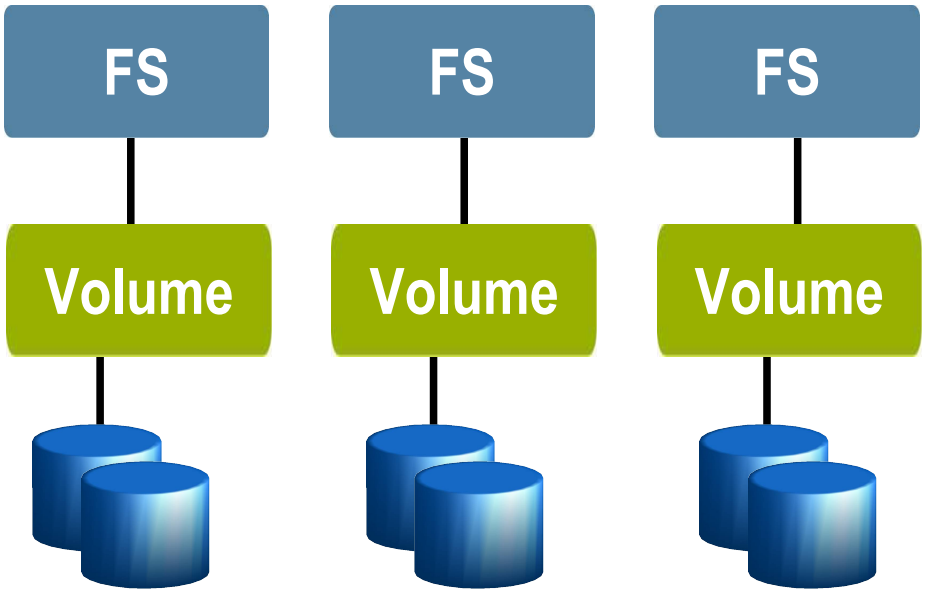
- Straightforward Commands and a GUI
  - > Snapshots & Clones
  - > Quotas & Reservations
  - > Compression
  - > Pool Migration
  - > ACLs for Security



FS/Volume Model vs. ZFS

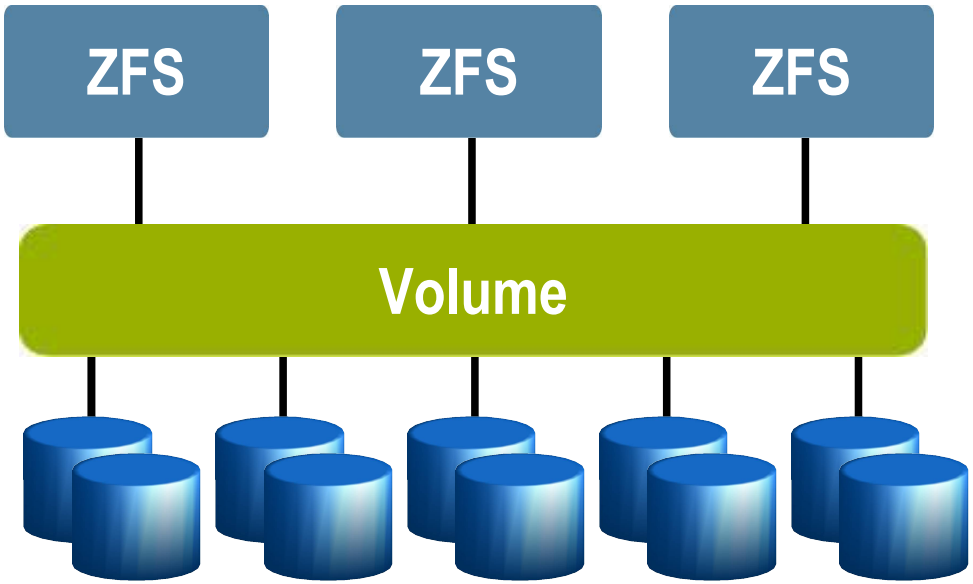
Traditional Volumes

- Abstraction: virtual disk
- Partition/volume for each FS
- Grow/shrink by hand
- Each FS has limited bandwidth
- Storage is fragmented, stranded



ZFS Pooled Storage

- Abstraction: malloc/free
- No partitions to manage
- Grow/shrink automatically
- All bandwidth always available
- All storage in the pool is shared



Agenda

- Introduction
- The Sun Fire X4500 Server and Use Cases
- Solaris 10 ZFS
- Summary

Sun Fire X4500 Server with Solaris 10 ZFS

- The First and Only Data Server
- Best Server Data Throughput and Storage Density
- Standard platform and common systems management capabilities

Sun Fire X4500 und Solaris ZFS - Neue Lösungen für große Datenmengen

Joachim Krebs
joachim.krebs@sun.com